A general sampling formula for community structure data

Haegeman, Bart; Etienne, Rampal S.

Published in:
Methods in ecology and evolution

DOI:
10.1111/2041-210X.12807

Publication date:
2017

Methods in Ecology and Evolution

BRITISH ECOLOGICAL SOCIETY

Methods in Ecology and Evolution 2017, **8,** 1506–1519 doi: 10.1111/2041-210X.12807

A general sampling formula for community structure data

Bart Haegeman*,[1] and Rampal S. Etienne[2]

[1]*Centre for Biodiversity Theory and Modelling, Theoretical and Experimental Ecology Station, CNRS and Paul Sabatier University, 2 route du CNRS, 09200 Moulis, France; and* [2]*Groningen Institute for Evolutionary Life Sciences, University of Groningen, Box 11103, 9700 CC Groningen, The Netherlands*

Summary

1. The development of neutral community theory has shown that the assumption of species neutrality, although implausible on the level of individual species, can lead to reasonable predictions on the community level. While Hubbell's neutral model and several of its variants have been analysed in quite some detail, the comparison of theoretical predictions with empirical abundance data is often hindered by technical problems. Only for a few models the exact solution of the stationary abundance distribution is known and sufficiently simple to be applied to data. For other models, approximate solutions have been proposed, but their accuracy is questionable.

2. Here, we argue that many of these technical problems can be overcome by replacing the assumption of constant community size (the zero-sum constraint) by the assumption of independent species abundances.

3. We present a general sampling formula for community abundance data under this assumption. We show that for the few models for which an exact solution with zero-sum constraint is known, our independent species approach leads to very similar parameter estimates as the zero-sum models, for six frequently studied tropical forest community samples.

4. We show that our general sampling formula can be easily confronted to a much wider range of datasets (very large datasets, relative abundance data, presence-absence data, and sets of multiple samples) for a large class of models, including non-neutral ones. We provide an R package, called SADISA (Species Abundance Distributions under the Independent Species Assumption), to facilitate the use of the sampling formula.

Key-words: density dependence, independent species, local community, metacommunity, multiple samples, neutral community model, presence-absence data, relative abundance, speciation model, species abundance distribution

Introduction

Species abundance distributions (SADs) have long intrigued ecologists (Fisher, Corbet & Williams 1943; Preston 1948; MacArthur 1957). The motivation is, besides the relative ease of collecting this type of data, that they may contain information on how species assemble in ecological communities, and on differences in species' properties. Indeed, intuitively a high abundance seems a sign of strong adaptation to the habitat where the species resides, indicating competitive dominance. However, such a high abundance perhaps just arises by chance. In the search for explanatory mechanisms, a plethora of models have been proposed to describe the SADs (McGill *et al.* 2007).

The last decade has seen a revived interest in the SAD because it is one of the key predictions of the neutral theory of biodiversity (Hubbell 2001; Rosindell, Hubbell & Etienne 2011), a theory that assumes that all individuals are functionally equivalent, regardless of the species it belongs to. This model attributes the differences in abundance not to differences in adaptation, but to inherent demographic stochasticity, i.e. a large abundance need not be a sign of strong adaptation, but is just due to demographic fortune. Comparing the neutral model predictions to those of more traditional niche-based models on abundance data has led to mixed results (Purves & Pacala 2005; Du, Zhou & Etienne 2011; Haegeman & Etienne 2011). This has invigorated the criticism that SADs do not contain sufficient information to infer the underlying process. However, stronger inferences might be possible when increasing the size of the community samples (Al Hammal *et al.* 2015). Moreover, in combination with other community patterns such as species-area curves, SADs may be informative (May, Huth & Wiegand 2015). Hence, it remains a useful exercise to fit reasonable models to species abundance data.

The central ingredient of fitting community models to data are sampling formulas. These formulas are used to evaluate the likelihood of data for a set of model parameters, find the optimal parameters using maximum likelihood and compare the fit quality of competing models, e.g. using Akaike information criterion. For Hubbell's neutral model, an exact sampling formula was derived by Etienne (2005). This formula gives the likelihood of observing S species abundances $n_1,n_2,\ldots,n_S$ in a sample of size J individuals according to a neutral model of a local community connected by immigration (described by the dispersal probability m, or equivalently by the dispersal number I) to a metacommunity governed by point-mutation

*Correspondence author. E-mail: bart.haegeman@sete.cnrs.fr

speciation (described by parameter θ, called the biodiversity number). However, this sampling formula is computationally demanding for samples of large size.

Nevertheless, the formula paved the way for a more general sampling theory (Etienne & Alonso 2005; Green & Plotkin 2007) in which the sampling formula was presented as a compound distribution of local, dispersal-limited sampling, and a metacommunity abundance distribution. It has been extended to multiple samples connected to the same metacommuity (Munoz *et al*. 2007; Etienne 2007, 2009), random-fission speciation (Haegeman & Etienne 2010; Etienne & Haegeman 2011) and multiple guilds (Janzen, Haegeman & Etienne 2015; see also Walker 2007). In all cases, the sampling formula was cumbersome to derive and demanding to compute and the total sample size allowing numerical computation was limited. Harris *et al*. (2017) circumvented the latter problem, but their approach is based on Bayesian computation rather than on a simple likelihood formula.

Here we present a new framework within which sampling formulas can be relatively easily derived and computed, not only for the models for which a zero-sum sampling formula is already available, but also for a wealth of other models. The crucial step is that we abandon the assumption of zero-sum dynamics, i.e. constant community size, and embrace the independent species assumption, i.e. we assume that species fluctuate independently of one another. It has been shown before that the zero-sum and independent species variants of neutral community models are intimately linked (Etienne, Alonso & McKane 2007a; Haegeman & Etienne 2008). In particular, the two model variants yield identical predictions for the local community model with fixed species pool and for the metacommunity model with point-mutation speciation. For Hubbell's neutral model, in which the local community model is coupled to the metacommunity model, the equivalence breaks down (Haegeman & Etienne 2011), but we show that there is still an excellent agreement, especially for highly diverse systems. We exploit this correspondence to derive sampling formulas that are easy to evaluate, even for very large sample size.

Independent-species approaches have been repeatedly applied to analyse the predictions of neutral community models. Alonso & McKane (2004) and Volkov *et al*. (2003, 2005, 2007) used this assumption to construct approximate solutions of the point-mutation speciation model. Haegeman & Etienne (2010) and Etienne & Haegeman (2011) used it as a starting point to get to a zero-sum sampling formula for random-fission speciation. Chisholm & Pacala (2010) and Haegeman & Etienne (2011) used it as a basis for a niche model. However, none of these studies have constructed a general framework to fit community models to abundance data, as we present here.

We start by providing an intuitive idea of the independent species approach and of its computational advantages over the standard zero-sum approach. Then, we present the general sampling formulas under the independent species assumption. We apply these formulas to the few models for which the zero-sum approach has been developed, and show that the independent species approach leads to very similar parameter estimates. Next, we present several model fitting problems which cannot be dealt with in the zero-sum framework, but for which the independent-species framework can be used. In particular, we consider community models with protracted speciation, species-level density dependence, and species-specific dispersal rates, and datasets of very large size, relative abundance data, presence-absence data and sets of multiple samples. In each of these cases the independent species framework leads to a straightforward fitting procedure, illustrating its simplicity and versatility. We provide an R package called SADISA (Species Abundance Distributions under the Independent Species Assumption) to evaluate the new sampling formulas.

From the zero-sum to the independent species assumption

The large majority of neutral community models is based on the zero-sum assumption. This assumption states that the number of individuals in the community is constant over time, implying that species abundance fluctuations are correlated: a decrease in one species has to be instantaneously compensated by an increase in another species. Here we explore the consequences of replacing the zero-sum by the independent species assumption, stating that species abundances fluctuate independently.

We illustrate the two assumptions using a simple community model. We consider a pool of species, whose relative abundances are assumed to be known and invariant over time (note that this assumption is limited to this example model; in the rest of the paper the species pool is governed by the probability distribution dictated by the metacommunity model). The dynamics of the local community coupled to this species pool consist of two processes: local mortality and immigration from the species pool (that is, we discard local reproduction; in the framework of Hubbell's model, this corresponds to setting $m = 1$ or $I \to \infty$; again, this assumption is limited to this example model). This holds for both the zero-sum and the independent species model variant of the model. The difference between the model variants resides in the way death and immigration events alternate. In the zero-sum version, each death event is immediately followed by an immigration event. As a result, the sum of all species abundance changes is zero (hence the term 'zero sum') and local community size remains constant over time. In the independent species version, each event, whether it is a death or an immigration, is uncoupled from other events. Hence, it is possible that several immigrations occur without any death in between them, or vice versa, so that the local community size would increase or decrease. In stationary state, however, the number of immigrations and deaths occurring over a longer period of time balance each other, so that the community size fluctuates around an average value. Moreover, because these stationary fluctuations are induced by independent events, the variability of community size is typically small. This strongly suggests that the predictions of the independent species model are often close to those of the zero-sum model. This is indeed what we find, as shown below.

In this paper we exploit the near equivalence of the two assumptions to simplify the evaluation of their model

predictions. Here we provide a first intuition of how this simplification works, while we refer to the next section for more details. We consider the case in which the species pool abundances are not known (if they are known, the evaluation of the zero-sum and independent species predictions are both straightforward). In this case, a community model at the regional scale (i.e. a metacommunity model) predicts the distribution of species pool abundances. We obtain the predictions for the local community abundances by averaging the local community composition for a given species pool over the distribution of species pool abundances. Under the zero-sum assumption, the species pool abundances are linked, and the computation of the average requires the evaluation of an S-dimensional integral, with S the number of species in the species pool. This is usually an extremely difficult numerical problem. In contrast, under the independent species assumption, species independence allows us to consider the S species one by one. As a result, the local community predictions decompose into S single-species averages, each of which requires the evaluation of a one-dimensional integral. This is an easy task, because the numerical integration of one-dimensional functions is not costly, even if there are many of them. Hence, by replacing the zero-sum by the independent species assumption, the evaluation of the model predictions simplifies drastically.

General sampling formula under the independent-species assumption

As for the zero-sum case, sampling formulas are the central ingredient of the inference procedure in the independent species case. These formulas give the probability of observing a specific set of abundance data under a community model for a specific set of parameters. Here we show that under the independent species assumption general sampling formulas can be derived, in contrast to the zero-sum assumption. Concrete examples for which independent species but not zero-sum formulas can be calculated are presented afterwards.

SINGLE-SAMPLE SAMPLING FORMULA

We first analyse the case in which a single sample taken from the community is available. We assume that the abundances of the species observed in the sample are quantified (in contrast to, e.g. presence-absence data). We represent the data as species abundance frequencies s_k, i.e. the number of species that are observed k times in the sample. For example, if there are nine observed species in the sample with abundances (species are ordered from most to least abundant),

Species #	1	2	3	4	5	6	7	8	9
Abundance in sample	11	5	5	4	2	1	1	1	1

then the corresponding abundance frequencies are $s_{11} = 1$, $s_5 = 2$, $s_4 = 1$, $s_2 = 1$, $s_1 = 4$, and all other $s_k = 0$.

Many independent species models have abundance frequencies that are approximately Poisson distributed. In Appendix S1, Supporting Information, we show that if the number of species in the metacommunity is Poisson distributed, the Poisson distribution is exact. Moreover, we argue that even if this condition is not met, the Poisson approximation is often very accurate. In those cases, which include all the independent species models considered in this paper, the independent species sampling formula is, either exactly or to a very good approximation, a product of Poisson samples,

$$\mathbb{P}(\mathcal{D}) = \prod_{k>0} \mathrm{e}^{-\lambda_k} \frac{\lambda_k^{s_k}}{s_k!}, \qquad \text{eqn 1}$$

where $\mathcal{D}$ stands for the data, i.e. the observed abundance frequencies. The numbers λ_k denote the predicted abundance frequencies, given by,

$$\lambda_k = \mathbb{E}s_k = \int \mathbb{P}(k|x)\rho(x)\mathrm{d}x. \qquad \text{eqn 2}$$

The term $\mathbb{P}(k|x)$ in the integrand of eqn (2) stands for the probability that a species with relative abundance x in the metacommunity is observed k times in the sample taken from the local community. For example, for neutral dispersal-limited sampling, it is given by a negative binomial distribution,

$$\mathbb{P}(k|x) = \frac{(Ix)_k(1-q)^{Ix}q^k}{k!}, \qquad \text{eqn 3}$$

with I the dispersal number and q a parameter that can be interpreted as sampling effort (see Appendix S2). The term $\rho(x)$ in the integrand of eqn (2) denotes the metacommunity abundance density, that is, $\rho(x)\mathrm{d}x$ gives the number of species with relative abundance in the interval $[x,x+\mathrm{d}x]$ in the metacommunity. For example, for a neutral model with point-mutation speciation, we have

$$\rho(x) = \theta\frac{\mathrm{e}^{-\theta x}}{x}, \qquad \text{eqn 4}$$

where θ is the metacommunity diversity (see Appendix S3). Note the similarity in model structure between local community and metacommunity: while the sum $\Sigma_{k=k_1}^{k_2}\lambda_k$ equals the expected number of species with abundance k between k_1 and k_2 in the local community, the integral $\int_{x_1}^{x_2}\rho(x)\mathrm{d}x$ equals the expected number of species with abundance x between x_1 and x_2 in the metacommunity. Also, the interpretation of variable x as relative abundance requires some care (see Appendix S3). The sum of x over all metacommunity species is equal to one only on average, although its fluctuations are often limited. Alternatively, variable x can be interpreted as an immigration propensity (see Appendix S3).

The evaluation of sampling formula (1) boils down to the computation of several integrals (2). It suffices to compute integrals λ_k for abundances k that are observed in the sample, i.e. for which $s_k > 0$. This can be seen by rewriting eqn (1) as

$$\mathbb{P}(\mathcal{D}) = \mathrm{e}^{-\Lambda} \prod_{k|s_k>0} \frac{\lambda_k^{s_k}}{s_k!} \qquad \text{eqn 5}$$

with Λ the expected number of observed species,

$$\Lambda = \sum_{k>0} \mathbb{E}s_k = \int \mathbb{P}(\text{obs}|x)\rho(x)\mathrm{d}x, \qquad \text{eqn 6}$$

where $\mathbb{P}(\text{obs}|x)$ is the probability that a species with relative abundance x in the metacommunity is present in the data, $\mathbb{P}(\text{obs}|x) = 1 - \mathbb{P}(0|x)$.

By substituting eqns (3) and (4) into eqns (2) and (1), we obtain a concrete sampling formula with model parameters θ, I and q. This formula can be directly used for likelihood maximization, and connects model predictions and empirical data. Regarding its application, the independent species sampling formula is very similar to the zero-sum sampling formula.

In comparison with the zero-sum case, the independent species sampling formula depends on an additional parameter, the sampling effort q. It is a number between 0 and 1; the larger this number, the larger the expected sample size (see Appendix S2). It can be estimated from the data, as the other model parameters. Alternatively, it can be determined a priori, based on the sample size J. The latter approach leads to a close correspondence with the zero-sum estimation procedure, in which the sample size J is also set beforehand. The parameter q can be tuned such that the expected sample size in the independent species approach matches the real sample size, which is also the fixed sample size used in the zero-sum approach. By applying this tuning, we obtain parameter estimates with the independent species approach that are almost identical to those obtained with the zero-sum approach, as we will show in the next section.

For the case of dispersal-limited sampling, given by eqn (3), the same sampling formula applies for the entire local community or for a sample taken from the local community. This is due to a property called sampling invariance (see Appendix S2). It suffices to set the parameter q in accordance with the size of the dataset, whether it is an exhaustive census or a non-exhaustive sample. In particular, the sampling formula does not depend on the size of the local community from which the sample was taken. However, sampling invariance, and the associated flexibility in dealing with either census or sample data, does not hold generally, as we will illustrate in the next section.

MULTIPLE-SAMPLES SAMPLING FORMULA

We now extend the sampling formula to L local communities connected to a single metacommunity. There is no direct migration between local communities; they are interdependent due to the immigration from the common metacommunity. We assume that we have a sample with abundance data taken from each of the local communities. As for the single-sample case, we express the data in terms of abundance frequencies. In particular, for each of the species observed in at least one of the L samples, we introduce the abundance vector $\vec{k} = (k_1, k_2, \ldots, k_L)$ containing its abundance in each sample. Abundance frequency $s_{\vec{k}}$ is equal to the number of species with abundance vector $\vec{k}$. For example, consider $L = 2$ local communities and suppose there are 8 observed species in total. If their abundances are given by,

Species #	1	2	3	4	5	6	7	8
Abundance in sample of 1st community	7	4	2	2	1	1	0	0
Abundance in sample of 2nd community	9	3	1	1	1	0	1	1

then the corresponding abundance frequencies are $s_{(7,9)} = 1$, $s_{(4,3)} = 1$, $s_{(2,1)} = 2$, $s_{(1,1)} = 1$, $s_{(1,0)} = 1$, $s_{(0,1)} = 2$, and all other $s_{\vec{k}} = 0$.

For independent species models the abundance frequencies are Poisson distributed, approximately if not exactly (see Appendix S1). The independent species sampling formula is

$$\mathbb{P}(\mathcal{D}) = \mathrm{e}^{-\Lambda} \prod_{\vec{k}|s_{\vec{k}}>0} \frac{\lambda_{\vec{k}}^{s_{\vec{k}}}}{s_{\vec{k}}!}, \qquad \text{eqn 7}$$

where $\lambda_{\vec{k}}$ is given by

$$\lambda_{\vec{k}} = \mathbb{E}s_{\vec{k}} = \int \left(\prod_{\ell=1}^{L} \mathbb{P}_\ell(k_\ell|x) \right) \rho(x)\mathrm{d}x, \qquad \text{eqn 8}$$

and Λ is given by

$$\Lambda = \sum_{\vec{k}|\sum_\ell k_\ell > 0} \int \mathbb{P}(\text{obs}|x)\rho(x)\mathrm{d}x. \qquad \text{eqn 9}$$

In these eqns $\mathbb{P}_\ell(k_\ell|x)$ is the probability of observing a species with relative abundance x in the metacommunity k_ℓ times in the sample taken from local community ℓ, and $\mathbb{P}(\text{obs}|x)$ is the probability of observing a species with relative abundance x in the metacommunity in at least one of the samples, i.e. $\mathbb{P}(\text{obs}|x) = 1 - \prod_\ell \mathbb{P}_\ell(0|x)$. For example, under neutral dispersal-limited sampling with dispersal number I_ℓ and sampling effort q_ℓ in the local community ℓ, we have

$$\mathbb{P}_\ell(k_\ell|x) = \frac{(I_\ell x)_{k_\ell}(1-q_\ell)^{I_\ell x} q_\ell^{k_\ell}}{k_\ell!}. \qquad \text{eqn 10}$$

Combining this expression with a choice for the metacommunity abundance density $\rho(x)$, we obtain a complete multiple-samples sampling formula.

MULTIPLE-GUILDS SAMPLING FORMULA

Another extension of the sampling formula consists in allowing for guild structure within the community (or communities) under study. We denote the number of guild by G, and we assume that they do not interact at the metacommunity level. The local community is composed of species that immigrated from the guild metacommunities, and the sample data is taken from the local community, possibly containing species of different guilds. We specify the data using abundance frequencies $s_k^{(g)}$, which are the number of species with abundance k in guild g. For example, if there are $G = 2$ guilds with species abundances,

	1st guild			2nd guild				
Species #	1	2	3	1	2	3	4	5
Abundance in sample	7	1	1	5	2	2	1	1

then $s_7^{(1)} = 1$, $s_1^{(1)} = 2$, $s_5^{(2)} = 1$, $s_2^{(2)} = 2$, $s_1^{(2)} = 2$, and all other $s_k^{(g)} = 0$.

The independent species sampling formula is, either exactly or approximately (see Appendix S1),

$$\mathbb{P}(\mathcal{D}) = \prod_{g=1}^{G}\left(\mathrm{e}^{-\Lambda^{(g)}}\prod_k \frac{\left(\lambda_k^{(g)}\right)^{s_k^{(g)}}}{s_k^{(g)}!}\right), \qquad \text{eqn 11}$$

where $\lambda_k^{(g)}$ and $\Lambda^{(g)}$ are given by eqns (2) and (6). Local sampling probabilities $\mathbb{P}^{(g)}(k|x)$ and metacommunity abundance densities $\rho^{(g)}(x)$ can be guild-dependent. Despite this complexity, sampling formula (11) expresses independence between species belonging to the same and to different guilds.

Comparison to models with zero-sum sampling formula

We compare the parameter estimates and likelihoods obtained with the independent species approach and the zero-sum approach, in those cases where a zero-sum sampling formula is available and computable.

SINGLE SAMPLES

The most studied neutral community model, also known as Hubbell's model, combines point-mutation speciation and dispersal-limited sampling (Hubbell 2001). To evaluate the zero-sum sampling formula, we follow the approach of Etienne (2005). This involves an arbitrary-precision computation with Stirling numbers, using the computer algebra system PARI/GP. The evaluation of the independent species sampling formula, given by eqns (1–4), requires the computation of several one-dimensional integrals. Because the integrands are often sharply peaked, we use a dedicated numerical integration algorithm, which is included in the R package SADISA.

We apply both sampling formulas to six datasets of tropical tree communities (Volkov *et al.* 2005; Etienne & Haegeman 2011). The parameter estimates obtained with the zero-sum and the independent species approach are very similar (Table 1, rows ZSC and ISA). Importantly, the likelihood values should not be compared, because they are not likelihoods for exactly the same data. The zero-sum approach assumes that the total number of individuals is given by the observed value, while the independent species approach treats this as additional data the probability of which is incorporated in the total likelihood. This explains why the zero-sum likelihood is systematically higher than the independent species likelihood (the log-likelihood is less negative, see Table 1). However, after conditioning the independent species likelihood on sample size (see Appendix S4), the zero-sum and independent species likelihood values almost coincide (Table 1, rows ZSC and ISAC). Note that the parameter estimates are even closer than in the case without conditioning (except for the Sinharaja dataset).

Table 1. Fits for neutral model with point-mutation speciation and dispersal-limited sampling. We analysed six datasets of tropical tree communities (Volkov *et al.* 2005; Etienne *et al.* 2007b; Etienne & Haegeman 2011), and we computed the maximum-likelihood fits for three model variants. The first variant, ZSC, imposes the zero-sum constraint, so that community size is invariant over time (results taken from Etienne *et al.* 2007b). The second variant, ISA, assumes independence between species. The third variant, ISAC, is also based on species independence, but the abundance distribution is conditioned on sample size. Note that likelihoods of model variants ZSC and ISAC are comparable (but the likelihood of ISA is not comparable with those of ZSC and ISAC)

Dataset	Model	θ	I	m	LL
BCI	ZSC	47·67	2211	0·0934	−308·73
	ISA	47·94	2175	0·0920	−317·70
	ISAC	47·67	2213	0·0935	−308·73
Korup	ZSC	52·73	29 700	0·5470	−317·04
	ISA	52·88	29 290	0·5436	−326·09
	ISAC	52·73	29 700	0·5471	−317·04
Pasoh	ZSC	190·9	2708	0·0926	−359·38
	ISA	191·4	2689	0·0919	−367·90
	ISAC	190·9	2712	0·0927	−359·38
Sinharaja	ZSC	436·8	32·38	0·0019	−252·93
	ISA	439·8	32·45	0·0019	−262·00
	ISAC	461·5	31·96	0·0019	−253·05
Yasuni	ZSC	204·2	13 170	0·4288	−297·15
	ISA	204·4	13 110	0·4277	−305·20
	ISAC	204·2	13 180	0·4289	−297·15
Lambir	ZSC	285·6	4296	0·1146	−386·38
	ISA	286·0	4280	0·1143	−394·93
	ISAC	285·5	4299	0·1147	−386·39

The likelihood landscapes for the zero-sum and the independent species approach are almost identical (Fig. 1). The ridge of high likelihood, present in both cases, is related to a well-known problem of Hubbell's neutral model, namely, the difficulty of distinguishing abundance distributions resulting from high regional diversity and low dispersal from those resulting from low regional diversity and high dispersal (Etienne *et al.* 2006). Clearly, the independent species approach has the same problem. Note that the colour code in the two panels is not exactly the same; the colour codes for the log-likelihood function differ by an additive constant. However, this constant difference has no effect on the maximum-likelihood estimates. Figure 2 shows that also the fitted SADs are almost identical. Hence, at least for the community model and the datasets considered here, the zero-sum approach and the independent species approach give practically equivalent results.

For two other speciation models, the zero-sum sampling formula for a single sample and single guild has been derived, assuming neutral dispersal-limited sampling. For random-fission speciation, the metacommunity abundance density $\rho(x)$ is given by (see Appendix S3; compare with eqn (4)),

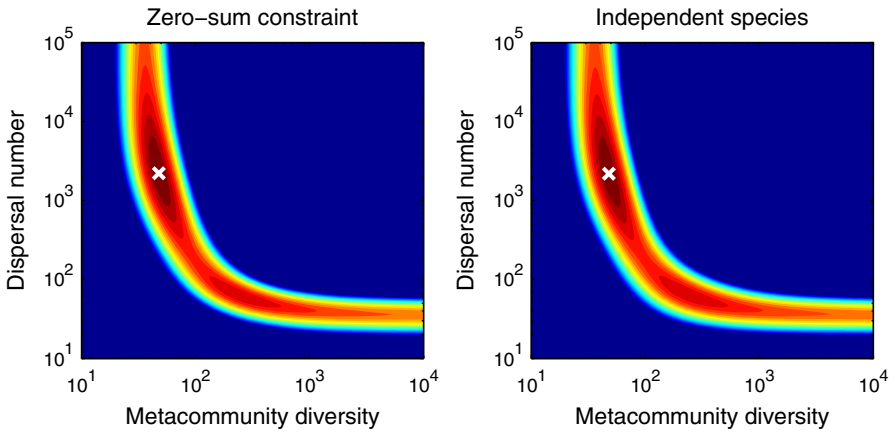

Fig. 1. Likelihood landscape for zero-sum and independent species approach. We consider the point-mutation speciation model with dispersal-limited sampling. We computed the zero-sum and independent-species likelihood as a function of metacommunity diversity θ (x-axis) and dispersal number I (y-axis) for the BCI dataset. Warmer colours correspond to higher likelihood values. The white $\times$-mark indicates the maximum-likelihood parameters. The two likelihood functions are almost identical, up to a constant factor (the colour code is relative to the maximum log-likelihood value; for example, dark blue corresponds to log-likelihood values at least 40 units below the maximum).

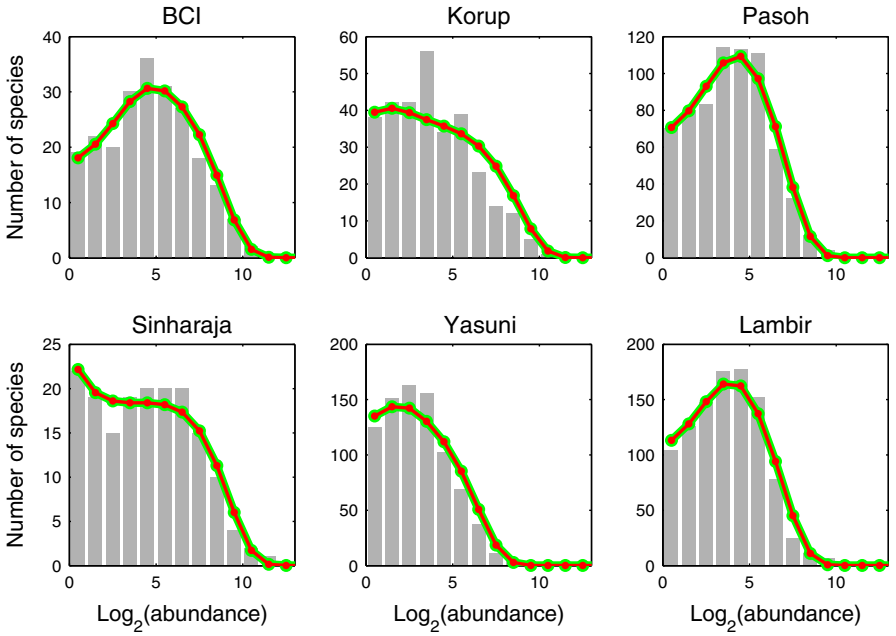

Fig. 2. Species abundance distributions for neutral model with point-mutation speciation and dispersal-limited sampling. For the six tropical forest plots (data represented by grey bars) we plot the fitted distributions with the zero-sum approach (thick green line) and the independent species approach (thin red line). The two fitted distributions are almost identical.

$$\rho(x) = \phi^2 e^{-\phi x}. \qquad \text{eqn 12}$$

Like θ for point mutation, the parameter ϕ characterizes the metacommunity diversity (in particular, it gives the expected number of species in the metacommunity). Also a model with per-species speciation has a zero-sum sampling formula (Etienne *et al.* 2007b). In the independent species setting, the metacommunity abundance density $\rho(x)$ is given by

$$\rho(x) = \frac{\theta^{1-\alpha}}{\Gamma(1-\alpha)} \frac{e^{-\theta x}}{x^{1+\alpha}}. \qquad \text{eqn 13}$$

Parameter θ is related to the per-individual speciation rate, while parameter α measures the importance of per-species speciation (with $0 \leq \alpha < 1$). The metacommunity diversity increases both with increasing θ and increasing α. Note that we recover the point-mutation model for $\alpha = 0$ and the random-fission model for $\alpha = -1$ (formally, because $\alpha = -1$ is outside the range $0 \leq \alpha < 1$ of values allowed by the per-species speciation model). While we do not have a direct independent species derivation of eqn (13), we show in Appendix S5 that this equation is the independent species equivalent of the zero-sum solution.

Table 2. Fits for neutral model with random-fission speciation and dispersal-limited sampling. Same datasets as in Table 1. We consider two model variants: variant ZSC imposes the zero-sum constraint (results taken from Etienne & Haegeman 2011); variant ISA assumes independence between species. ZSC and ISA likelihoods are not comparable. In column ΔLL we compare the maximum log-likelihoods of the random-fission model with those of the point-mutation model, for the ZSC and the ISA variant

Dataset	Model	ϕ	I	m	LL	ΔLL
BCI	ZSC	595·1	61·61	0·0029	−311·92	−3·20
	ISA	595·2	61·81	0·0029	−321·11	−3·41
Korup	ZSC	∞	49·52	0·0020	−318·67	−1·63
	ISA	∞	49·61	0·0020	−327·75	−1·66
Pasoh	ZSC	1528	263·4	0·0098	−363·75	−4·37
	ISA	1527	264·0	0·0098	−372·49	−4·58
Sinharaja	ZSC	927·6	32·42	0·0019	−252·88	+0·05
	ISA	950·1	32·35	0·0019	−261·97	+0·03
Yasuni	ZSC	10 980	197·0	0·0111	−306·75	−9·60
	ISA	11 130	196·9	0·0111	−314·88	−9·68
Lambir	ZSC	2500	372·5	0·0111	−402·32	−15·94
	ISA	2500	372·9	0·0111	−411·08	−16·15

Similarly to the case of point mutation, we find that the zero-sum and independent species estimates are very close, both for the random-fission speciation model (Table 2) and for the per-species speciation model (Table 3). The absolute log-likelihood values should not be compared (because they are not likelihoods for exactly the same data, see above), but the log-likelihood values relative to the point-mutation values are comparable. The log-likelihood differences ΔLL are very similar in all cases, showing that the zero-sum approach and the independent species approach lead to the same inferences.

The independent species sampling formula (1) is only approximately valid for these two speciation models (see Appendix S1). Nevertheless, the agreement with the zero-sum results is as strong as for the case of point-mutation speciation, for which the independent species sampling formula (1) is exact. This indicates, in addition to the general argument of Appendix S1, that the Poisson approximation is very accurate.

The data provides stronger support for point-mutation speciation than for random-fission speciation, as reported by Etienne & Haegeman (2011). The data does not contain signs of per-species speciation in the case without dispersal limitation, in agreement with Etienne *et al.* (2007b). However, in the case with dispersal limitation, which has not been studied previously, there is strong evidence of per-species speciation in the Korup and Yasuni datasets. Hence, the selection between speciation models depends on whether or not dispersal limitation is taken into account. While this is an intriguing result, an analysis of its precise meaning is beyond the scope of this paper.

MULTIPLE SAMPLES

The zero-sum analog of the multiple-samples sampling formula (7) has only been explored for the point-mutation

Table 3. Fits for per-species speciation model, or equivalently, metacommunity model with density dependence. Same datasets as in Table 1. Model variants are combinations of nDL, no dispersal limitation; DL, dispersal limitation; ZSC, zero-sum constraint; ISA, species independence approach. Results for model (nDL, ZSC) are taken from Etienne *et al.* (2007b), but results for model (DL, ZSC) have not been reported before. The maximum likelihood of the per-species speciation model is always larger than the corresponding point-mutation likelihood (column ΔLL), because point-mutation speciation is a special case of per-species speciation (case $\alpha = 0$)

Dataset	Model		$\theta = \frac{\nu_0 + \nu_1 J_M}{1-\nu_1}$	$\alpha = \frac{\nu_0}{1-\nu_1}$	I	m	LL	ΔLL
BCI	nDL	ZSC	34·97	0	∞	1	−318·85	0
	nDL	ISA	35·06	0	∞	1	−327·97	0
	DL	ZSC	38·32	0·1203	1049	0·0466	−308·19	0·54
	DL	ISA	37·33	0·1354	960·2	0·0428	−317·01	0·69
Korup	nDL	ZSC	44·54	0·0289	∞	1	−318·31	0·36
	nDL	ISA	44·19	0·0303	∞	1	−327·35	0·40
	DL	ZSC	13·87	0·4326	1046	0·0408	−306·82	10·22
	DL	ISA	12·99	0·4420	996·8	0·0390	−315·38	10·71
Pasoh	nDL	ZSC	126·4	0	∞	1	−392·51	0
	nDL	ISA	126·7	0	∞	1	−401·20	0
	DL	ZSC	184·2	0·0361	2192	0·0763	−359·31	0·07
	DL	ISA	183·0	0·0447	2081	0·0727	−367·80	0·11
Sinharaja	nDL	ZSC	25·63	0	∞	1	−253·78	0
	nDL	ISA	25·73	0	∞	1	−262·82	0
	DL	ZSC	12·72	0·5123	145·3	0·0085	−252·13	1·19
	DL	ISA	11·77	0·5270	138·8	0·0081	−260·59	1·42
Yasuni	nDL	ZSC	178·3	0	∞	1	−307·58	0
	nDL	ISA	178·6	0	∞	1	−315·68	0
	DL	ZSC	61·86	0·5272	1117	0·0598	−278·88	18·27
	DL	ISA	60·39	0·5324	1098	0·0589	−286·54	18·66
Lambir	nDL	ZSC	195·0	0	∞	1	−437·89	0
	nDL	ISA	195·3	0	∞	1	−446·57	0
	DL	ZSC	245·5	0·1161	2546	0·0713	−385·20	1·18
	DL	ISA	244·3	0·1202	2503	0·0702	−393·65	1·28

Table 4. Fits for multiple samples. From the abundance data of three Panamian forest plots, we constructed eleven datasets, each consisting of three samples (one full dataset, and ten reduced datasets; see Etienne (2007) for details). We computed the maximum-likelihood fits for two model variants. The first variant, ZSC, imposes the zero-sum constraint (results taken from Etienne 2007). The second variant, ISA, assumes independence between species. Likelihoods of the two model variants are not comparable

Dataset	Model	θ	I	LL
Full dataset	ZSC	259·3	44·24	−1091·80
	ISA	259·4	44·46	−1116·12
Subsample 1	ZSC	270·5	39·18	−679·87
	ISA	270·8	39·41	−702·08
Subsample 2	ZSC	273·9	39·21	−668·84
	ISA	274·2	39·44	−690·96
Subsample 3	ZSC	280·0	41·18	−673·74
	ISA	280·2	41·41	−695·75
Subsample 4	ZSC	282·2	42·63	−680·40
	ISA	282·4	42·87	−702·35
Subsample 5	ZSC	290·8	41·71	−679·28
	ISA	291·1	41·94	−701·23
Subsample 6	ZSC	297·3	39·13	−654·40
	ISA	297·6	39·35	−676·45
Subsample 7	ZSC	298·6	37·27	−652·12
	ISA	299·0	37·48	−674·39
Subsample 8	ZSC	296·5	36·32	−640·46
	ISA	296·8	36·53	−662·70
Subsample 9	ZSC	300·4	37·65	−647·22
	ISA	300·7	37·87	−669·34
Subsample 10	ZSC	271·5	40·47	−688·08
	ISA	271·7	40·70	−710·15

speciation process and neutral dispersal-limited sampling (Etienne 2007; Connolly, Hughes & Bellwood 2017). Here we apply the independent species sampling formula (7) on the same datasets. We follow the approach of Etienne (2007) and reduce the number of parameters to estimate by assuming that $I_\ell = I$ for all ℓ. Moreover, we eliminate the sampling efforts q_ℓ by setting the expected sample size equal to the observed sample size for each local community ℓ. As a result, the likelihood has to be maximized over two parameters only (θ and I).

We find very good agreement between the estimates obtained with the zero-sum constraint and those obtained with the independent species assumption (Table 4). The likelihood values are different, but as explained before, they should not be compared. Indeed, the zero-sum approach imposes a constraint on the allowed datasets that is not present in the independent species approach.

MULTIPLE GUILDS

Recently, we derived the zero-sum sampling formula for a single sample of two dispersal guilds with a metacommunity governed by point-mutation speciation (Janzen, Haegeman & Etienne 2015). As we were interested in detecting guild differences in dispersal rate, we assumed that the two guilds have the same distribution of relative abundances in the metacommunity, but no species in common. Here we apply the multiple-guilds sampling formula (11) of the independent species approach to the dataset studied by Janzen, Haegeman & Etienne (2015).

Importantly, the assumption that the guild metacommunities do not differ can be implemented in different ways. The zero-sum approach of Janzen, Haegeman & Etienne (2015) assumed that the two guilds have the same speciation rates, and hence, the same metacommunity diversity θ (denoted by 'sS', which stands for same speciation rate). However, this assumption does not eliminate differences in guild metacommunity sizes. One can therefore impose additionally that guild metacommunity sizes are the same (denoted by 'sM', which stands for same metacommunity size). It turns out that this additional assumption has a strong effect on the parameter estimates [Table 5; compare rows (sM, ZSC) and (sS, ZSC)], regardless of whether guilds have the same or different dispersal rates: the likelihood is consistently higher for the second implementation (same speciation rate and same guild metacommunity size) than for the first implementation (same speciation rate, but guild metacommunity size can vary).

This distinction is crucial for the comparison of the zero-sum and independent species estimates. The independent species model underlying sampling formula (11) corresponds to the second implementation, i.e. the identity of guild speciation rates implies the identity of guild metacommunity sizes. Indeed, the independent species estimates are very similar to the zero-sum estimates obtained with the second implementation [Table 5; compare rows (sM, ZSC) and (sM, ISA)]. This agreement holds both when assuming that guilds have the same or different dispersal rates. Note that there is no independent species model that corresponds to the first implementation, where guild metacommunity sizes can vary.

Extensions to models without zero-sum sampling formula

We study several problems of fitting community models to abundance data for which the zero-sum approach does not lead to a workable solution. We show that by adapting the independent species approach each of these problems can be solved without major obstacles.

DIFFERENT $\mathbb{P}(k|x)$: LOCAL COMMUNITY MODELS

Until now we have assumed that the sampling probability is given by neutral dispersal-limited sampling (3). The independent species framework allows us to analyse other local community models. As an illustration, we consider a model with density dependence, which constitutes a departure from neutrality (see Allouche & Kadmon 2009; Jabot & Chave 2011 for other extensions of the neutral model with density dependence).

Many forms of density dependence can be incorporated in the independent species framework. We assume that the per capita birth rate is proportional to $1 - \frac{\alpha}{k}$ and that the per capita death rate is constant. This leads to positive density dependence for $0 < \alpha < 1$ and negative density dependence for $\alpha < 0$. In Appendix S6 we show that the sampling probability $\mathbb{P}(k|x)$ then becomes,

Table 5. Fits for multiple guilds. Guild 1: species with biotic dispersal; guild 2: species with abiotic dispersal; see Janzen, Haegeman & Etienne (2015) for details. For six censuses of the BCI plot we computed the maximum-likelihood fits for several model variants: sM, guild metacommunities have same size; sS, guilds have same speciation rate; dD, guilds have different dispersal rate; sD, guilds have same dispersal rate; ZSC, zero-sum constraint; ISA, species independence approach. Results for model (sS, ZSC) are taken from Janzen, Haegeman & Etienne (2015), but results for model (sM, ZSC) have not been reported before

Dataset	Model			θ	I_1	I_2	LL
BCI (1982)	sM	dD	ZSC	80·50	2433	13·56	−365·92
	sM	dD	ISA	80·85	2399	13·90	−382·59
	sM	sD	ZSC	41·22	79 520	79 520	−410·32
	sM	sD	ISA	41·49	71 420	71 420	−426·80
	sS	dD	ZSC	503·0	49·91	7·871	−368·06
	sS	sD	ZSC	67·29	520·7	520·7	−399·18
BCI (1985)	sM	dD	ZSC	79·43	2743	12·75	−365·39
	sM	dD	ISA	79·77	2704	13·08	−382·07
	sM	sD	ZSC	∞	20·31	20·31	−411·55
	sM	sD	ISA	∞	20·41	20·41	−428·05
	sS	dD	ZSC	561·0	47·76	7·338	−367·52
	sS	sD	ZSC	65·57	573·4	573·4	−400·82
BCI (1990)	sM	dD	ZSC	78·62	2078	12·52	−361·33
	sM	dD	ISA	78·92	2059	12·86	−378·08
	sM	sD	ZSC	42·19	8137	8137	−407·51
	sM	sD	ISA	42·53	7803	7803	−424·00
	sS	dD	ZSC	107·0	53·68	7·546	−365·42
	sS	sD	ZSC	62·13	583·7	583·7	−393·86
BCI (1995)	sM	dD	ZSC	77·93	2078	12·05	−371·03
	sM	dD	ISA	78·24	2057	12·37	−387·83
	sM	sD	ZSC	41·31	9329	9329	−417·96
	sM	sD	ISA	41·65	8859	8859	−434·49
	sS	dD	ZSC	106·5	53·32	7·277	−374·98
	sS	sD	ZSC	62·00	554·1	554·1	−404·08
BCI (2000)	sM	dD	ZSC	77·77	2060	12·53	−361·10
	sM	dD	ISA	78·08	2040	12·86	−377·85
	sM	sD	ZSC	42·08	7148	7148	−405·99
	sM	sD	ISA	42·41	6897	6897	−422·49
	sS	dD	ZSC	105·8	54·41	7·594	−364·99
	sS	sD	ZSC	61·12	595·6	595·6	−392·54
BCI (2005)	sM	dD	ZSC	76·09	2589	13·01	−359·54
	sM	dD	ISA	76·39	2558	13·37	−376·26
	sM	sD	ZSC	40·50	21 040	21 040	−401·50
	sM	sD	ISA	40·79	19 980	19 980	−417·99
	sS	dD	ZSC	471·3	48·09	7·665	−361·97
	sS	sD	ZSC	60·41	669·9	669·9	−390·99

$$\mathbb{P}(k|x) = \begin{cases} \frac{Ix(Ix-\alpha)_k}{(1-q)^{-Ix+\alpha}Ix-\alpha}\frac{q^k}{k!} & \text{if } k \geq 1 \\ \frac{Ix-\alpha}{(1-q)^{-Ix+\alpha}Ix-\alpha} & \text{if } k = 0. \end{cases} \quad \text{eqn 14}$$

This expression replaces eqn (3) in sampling formula (1). Note that the sampling formula with density dependence lacks sampling invariance, that is, eqn (14) changes when considering a sample taken from the local community rather than the entire local community. This implies that, when applied to sample abundance data, the sampling formula depends on local community size, introducing an additional parameter to estimate. When fitting the model to the tropical forest plots, we find some evidence of negative density dependence in the local community (Table S1).

DIFFERENT ρ(x): METACOMMUNITY MODELS

The metacommunity abundance density $\rho(x)$ depends on the metacommunity dynamics. Particular interest has been given to how new species arise. Rosindell *et al.* (2010) proposed the protracted speciation model to account for the fact that speciation takes time. In Appendix S3 we show that the corresponding metacommunity abundance density $\rho(x)$ is given by

$$\rho(x) = \theta\frac{e^{-\frac{\theta\phi}{\theta+\phi}x} - e^{-\phi x}}{x}. \quad \text{eqn 15}$$

Parameter θ is related to the speciation-initiation rate, while parameter φ is inversely proportional to speciation time. Interestingly, in the limit $\phi\to\infty$ we recover (4) for point-mutation speciation, and in the limit $\theta\to\infty$ we recover (12) for random-fission speciation. Hence, the protracted-speciation model interpolates between the two speciation models. Fitting the model to the six tropical forest plots shows that protractedness cannot be detected in the SADs (Table S2). Rosindell *et al.* (2010) reached the same conclusion using the approximate fitting procedure of Alonso & McKane (2004). Note that this procedure can be reinterpreted in the independent species framework (see Discussion).

As another example, we consider a metacommunity model with density dependence. Density dependence at large scales can effectively emerge from local interactions (Steele & Forrester 2005). We take the same form of density dependence as in the local community example: the per capita birth rate is proportional to $1 - \frac{\alpha}{k}$ and the per capita death rate is constant. The corresponding abundance density $\rho(x)$ is given by (see Appendix S5),

$$\rho(x) = \frac{\theta^{1-\alpha}}{\Gamma(1-\alpha)}\frac{e^{-\theta x}}{x^{1+\alpha}}, \quad \text{eqn 16}$$

which, interestingly, is the same expression as (13) for per-species speciation. However, where in the case of per-species speciation only positive values of α were meaningful (in particular, $0 \leq \alpha < 1$), the density-dependence interpretation of eqn (16) also allows negative values of α (in case of negative density dependence). The model fits for the tropical forest data have positive values of α (Table 3, rows DL). Hence, the interpretation is not univocal: it can indicate either per-species speciation or positive density dependence.

SPECIES-DEPENDENT PARAMETERS

The previous models are based on the assumption of species equivalence. While species differences are difficult to deal with in the zero-sum framework (Zhou & Zhang 2008), they can be easily incorporated with the independent species approach. Indeed, because the likelihood is equal to the product of species-level likelihoods, it suffices to introduce species-dependent parameters in each of the factors of this product. However, this

leads to likelihood functions of a large number of parameters (proportional to the number of species), which cannot be inferred from the data. To reduce the number of parameters, we consider an alternative model in which parameters differ between species, but species-specific parameters are drawn from a distribution that is the same for all species. Likelihood maximization can then be used to infer information about this distribution.

As an example, we suppose that dispersal number I differs between species and that the species-specific dispersal numbers I_i are drawn from distribution $\sigma(I)$. In Appendix S7 we show that the independent species sampling formula (1) still holds, with λ_k given by (instead of eqn 2),

$$\lambda_k = \int \mathbb{P}(k|x, I)\rho(x)\sigma(I)\mathrm{d}x\mathrm{d}I, \qquad \text{eqn 17}$$

and Λ given by (instead of eqn 6),

$$\Lambda = \int \mathbb{P}(\mathrm{obs}|x, I)\rho(x)\sigma(I)\mathrm{d}x\mathrm{d}I,$$

In a concrete application, one could parameterize the distribution $\sigma(I)$ by its variance, and infer this parameter from the data. If the likelihood for non-zero variance is higher than the likelihood for zero variance, there might be evidence that the dispersal number I differs between species. The strength of the evidence can be quantified, using likelihood-ratio tests. Note that this procedure informs us only on the existence of species differences in dispersal rate, but not on the dispersal rate of specific species.

A similar approach could be applied to other model parameters. For example, in the multiple-sample case, one could assume that dispersal number I differs between samples. To limit the number of parameters, i.e. to avoid the introduction of a parameter for each patch, one could assume that the sample-specific dispersal numbers I_ℓ are drawn from a common distribution $\sigma(I)$. The corresponding sampling formula can then be constructed along the lines explained above. However, because different species are affected by the same choice of dispersal number I_ℓ, the likelihood has no longer the product structure of independent species, so that the sampling formula is more complicated to evaluate.

LARGE DATASETS

Even if the zero-sum sampling formula is available, its evaluation becomes often cumbersome for large datasets. We have argued above that the independent species sampling formula is easier to evaluate. To further support this statement, we consider Hubbell's neutral model (point-mutation speciation and dispersal-limited sampling). For a fixed set of parameter values (metacommunity diversity $\theta = 50$ and dispersal number $I = 1000$), we generate sample data for sample sizes ranging from $J = 10^3$ to $J = 10^6$. This can be easily done within the independent species framework, because the abundance frequencies are independent Poisson random variables, see eqn (1). For each of the generated samples, we fit the model parameters, using maximum likelihood, once with the zero-sum sampling formula and once with the independent species sampling formula. We then compare the time it takes to complete the maximization. Note that one maximization typically requires a few hundreds of sampling formula evaluations.

The comparison results are shown in Fig. 3. The scaling of computation time with sample size differs between the two approaches: the independent species computation time scales as $\sqrt{J}$, and the zero-sum computation time scales as J^2. The independent species approach is faster for sample size $J > 10^4$. For example, for $J = 10^5$, the independent species computation takes about a minute, while the zero-sum computation takes about half an hour (on a standard laptop computer; see Fig. 3 for specifications). For still larger sample size, $J > 2 \times 10^5$, our implementation of the zero-sum computation does not complete, due to memory problems that occurred during the computation of large Stirling numbers (on which the zero-sum sampling formula is based; see Etienne 2005). In contrast, the independent species computation time remains below a few minutes for sample size J up to 10^6.

As an illustration, we fit Hubbell's model to an extended dataset of the BCI tropical forest plot, which includes all trees with dbh (diameter at breast height) above 1 cm (rather than trees with dbh above 10 cm). Due to the large sample size ($J \approx 2{\cdot}3 \times 10^5$), we were not able to evaluate the zero-sum likelihood on our computer. Likelihood maximization using the independent species approach did not pose any problem (see Table S3).

RELATIVE ABUNDANCE DATA

Another limitation of the zero-sum sampling formula is that it can only be applied to absolute species abundances. However, abundance data are often available as relative abundances (e.g. vegetation cover, biomass, fingerprint data). The independent species approach can be easily extended to that type of data. with sampling formula,

$$\mathbb{P}(\mathcal{D}) = \mathrm{e}^{-\Lambda}\prod_i \int_x \mathbb{P}(\mathrm{obs}|p_i)\mathbb{P}(p_i \in \mathrm{d}p_i|x)\rho(x)\mathrm{d}x, \qquad \text{eqn 18}$$

with p_i the observed relative abundance and Λ the expected number of observed species,

$$\Lambda = \int_p \int_x \mathbb{P}(\mathrm{obs}|p)\mathbb{P}(p \in \mathrm{d}p|x)\rho(x)\mathrm{d}x.$$

The integrand in eqn (18) contains two sampling probabilities. The first one is the probability density $\mathbb{P}(p \in \mathrm{d}p|x)$ for local relative abundance p given metacommunity relative abundance x. For the case of neutral dispersal-limited sampling, it is the continuous version of the negative binomial distribution (3), which is the gamma distribution,

$$\mathbb{P}(p \in \mathrm{d}p|x) = \frac{I^{Ix}}{\Gamma(Ix)}p^{Ix-1}\mathrm{e}^{-Ip}. \qquad \text{eqn 19}$$

The second one is the probability $\mathbb{P}(\mathrm{obs}|p)$ to observe in the sample a species with local relative abundance p. For example,

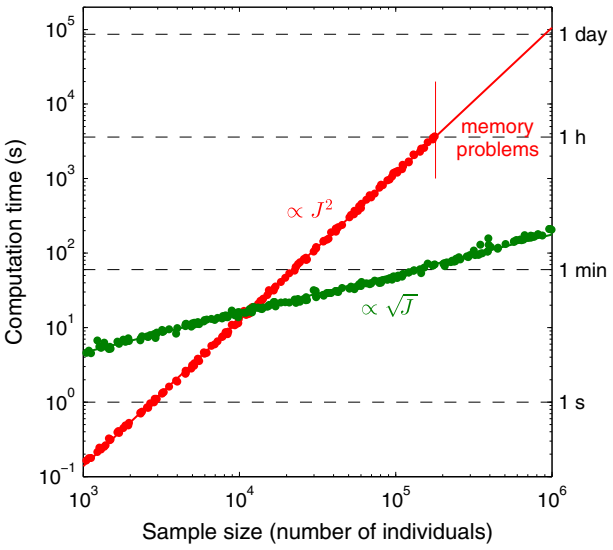

Fig. 3. Computational complexity of zero-sum and independent species likelihood maximization. We generated samples of different size for the neutral community model with point-mutation speciation ($\theta = 50$) and dispersal limitation ($I = 1000$), and estimated the model parameters, using the zero-sum (red dots) and independent species (green dots) sampling formula. Computation time scales consistently with sample size J: proportional to J^2 for the zero-sum approach (red line) and proportional to $\sqrt{J}$ for the independent species approach (green line). We did not succeed in evaluating the zero-sum likelihood for sample size $J > 2 \times 10^5$ due to memory problems (vertical red line). Computations were performed on a laptop computer with Intel Core i5 microprocessor (two cores, 2·80 GHz clock speed and 6 MB on-board memory) and 3·8 GB main memory.

one could take $\mathbb{P}(\text{obs}|p) = 1 - \mathrm{e}^{-\xi p}$, so that species with relative abundance under the threshold relative abundance $1/\xi$ are typically not detected, and species with relative abundances above it have a substantial chance of being detected. Note that sampling formula (18) can be generalized to multiple samples,

$$\mathbb{P}(\mathcal{D}) = \mathrm{e}^{-\Lambda} \prod_{i|\sum_\ell p_{i\ell}>0} \int_x \left(\prod_{\ell|p_{i\ell}>0} \mathbb{P}_\ell(\text{obs}|p_{i\ell}) \mathbb{P}_\ell(p_{i\ell} \in \mathrm{d}p_{i\ell}|x) \right) \left(\prod_{\ell|p_{i\ell}=0} \mathbb{P}_\ell(\text{unobs}|x) \right) \rho(x)\mathrm{d}x, \qquad \text{eqn 20}$$

with $\mathbb{P}_\ell(\text{unobs}|x) = 1 - \int_p \mathbb{P}_\ell(\text{obs}|p)\mathbb{P}(p \in \mathrm{d}p|x)$. The index i runs over all species that are observed at least in one sample. The index ℓ runs over the local communities from which a sample is taken; the first product inside the integrand corresponds to samples in which species i is observed, while the second product corresponds to samples in which species i is unobserved.

PRESENCE-ABSENCE DATA

We can apply our approach also to datasets where only species occurrences were scored in multiple sites, i.e. presence-absence data. We consider L samples. We introduce the presence-absence vector $\vec{o}$ of a species, i.e. $\vec{o} = (o_1, o_2, \ldots, o_L)$ with $o_\ell = 1$ if the species is present in sample ℓ and $o_\ell = 0$ if not. We denote the corresponding abundance frequencies by $s_{\vec{o}}$. Then, the independent species sampling formula is,

$$\mathbb{P}(\mathcal{D}) = \mathrm{e}^{-\Lambda} \prod_{\vec{o}} \frac{\lambda_{\vec{o}}^{s_{\vec{o}}}}{s_{\vec{o}}!}, \qquad \text{eqn 21}$$

with

$$\lambda_{\vec{o}} = \int \left(\prod_\ell \mathbb{P}_\ell(o_\ell|x) \right) \rho(x)\mathrm{d}x, \qquad \text{eqn 22}$$

and $\mathbb{P}_\ell(o_\ell = 1|x)$ the probability that a species with metacommunity abundance x is present in sample ℓ. For neutral dispersal-limited sampling (with dispersal number I_ℓ and sampling effort q_ℓ), we have (see eqn 10),

$$\mathbb{P}_\ell(o_\ell = 1|x) = \mathbb{P}_\ell(k_\ell \geq 1|x) = 1 - \mathbb{P}_\ell(k_\ell = 0|x) = 1 - (1 - q_\ell)^{I_\ell x}.$$

Discussion

We have provided a framework to compute, under the independent species assumption, a sampling formula for all mainland-island(s) models for which we can specify the metacommunity abundance density $\rho(x)$ and the local sampling probability $\mathbb{P}(k|x)$. The computational complexity of the sampling formula reduces to the evaluation of one-dimensional integrals of the form $\int \mathbb{P}(k|x)\rho(x)\mathrm{d}x$. Because the integrands are often sharply peaked, the numerical evaluation of these integrals can be challenging. We include a dedicated integration algorithm in the R package SADISA (which stands for Species Abundance Distributions under the Independent Species Assumption). Currently, the package implements the sampling formulas only for the analyses presented in the paper. However, it is relatively straightforward to use the methods implemented in the package for other community models.

The independent species framework allows us to fit a broad set of neutral community models. This set is much broader than the models with zero-sum sampling formulas, for which our approach is often (much) more efficient. The framework can be applied to larger datasets (higher abundances, more species, more samples) and to relative abundance and presence-absence data. The only requirement is the specification of the metacommunity abundance density $\rho(x)$ – which depends on the speciation process – and the local sampling probability $\mathbb{P}(k|x)$ – which depends on the local demographic dynamics. Even in cases where the independent species sampling formulas are approximate, such as the random-fission and the per-species speciation models, the parameter estimates are almost indistinguishable from the zero-sum results. The approach is not restricted to neutral scenarios, as illustrated by our examples of density dependence and species-dependent parameters. Independent-species models can be easily simulated, because the abundance frequencies are independent Poisson random variables (see Appendix S1). Simulated datasets are useful to explore

model predictions, but also to evaluate the accuracy of parameter estimates and the reliability of model inference (see below).

We have shown that the sampling formulas under the independent species assumption yield parameter estimates that are very similar to those obtained under the zero-sum constraint. This need not always be the case. The condition for this similarity is that the community size distribution is sharply peaked. This happens for the local community when the dispersal number I is large (e.g. $I > 10$; see Appendix S2), and in the metacommunity (under point mutation) when the diversity parameter θ is large (e.g. $\theta > 10$; see Appendix S3). Sampling formulas are typically applied to highly diverse systems, because only those systems are considered to contain sufficient information (i.e. enough 'replicates') to reliably estimate the parameters. Hence, we expect that the zero-sum and independent species fits will often agree. Even if the fits do not agree, this discrepancy should not be seen as a failure of the independent-species approach. Independent-species models are not only approximations of zero-sum models; they are fully consistent mathematical models in their own right. However, in such (rare) cases of discrepancy, the ecological meaning should be critically evaluated.

Our work sheds new light on previous attempts to link abundance data with community models. Alonso & McKane (2004) proposed a somewhat *ad hoc* approach to fit community models to abundance data. Within the independent-species framework, it corresponds to applying an additional conditioning on the observed number of species. As our approach does not have this conditioning, it does not discard the information contained in the observed number of species, and is thus more powerful. Volkov *et al.* (2003) combined the independent species metacommunity abundance density under point mutation with the zero-sum version of local dispersal-limited sampling. This mixed approach can be used to compute the expected abundance distribution, but is less helpful to derive the full sampling formula. We have shown how a consistent application of the independent species approach readily provides both the abundance distribution and the sampling formula. Green & Plotkin (2007) proposed abundance distributions which have the same structure as the ones we obtained from solving the independent species community models (compare their eqn 1 with our eqn 2). Our results can be interpreted as a more mechanistic underpinning of their distributions. Moreover, our framework indicates how to incorporate their abundance distributions into sampling formulas, which can then be used for parameter estimation and model selection.

The theory we have developed results in a long list of sampling formulas (see Appendix S8). The question arises how to choose among them in practice. The general structure of the sampling formula is dictated by the nature of the data: is the data expressed in absolute abundances, relative abundances, or as presence-absence data; is there a single or are there multiple samples? The biological question determines the different processes to include in the community models, which in turn determine the functions appearing in the sampling formula: the abundance density $\rho(x)$ at the regional scale, and the sampling probability $\mathbb{P}(k|x)$ at the local scale. We have presented a derivation for several of these functions, which can serve as a template for other community models. Once the functions $\rho(x)$ and $\mathbb{P}(k|x)$ have been specified, we can apply the independent species formalism to evaluate the sampling formula and to determine the maximum-likehood parameters. The R package SADISA includes a step-by-step demonstration for single-sample and multiple-samples examples.

Reliable inference of community processes from abundance data is well-known to be very challenging. While the independent species approach drastically simplifies the evaluation of the likelihood function, it evidently does not resolve fundamental issues of fitting community models to abundance data. For example, in Hubbell's neutral model, very large samples are required to distinguish between cases with high regional diversity and low dispersal and cases with low regional diversity and high dispersal (see the ridge of high likelihood in Fig. 1). Community structure is the result of the interplay between several processes, both at local and regional scales, which are often difficult to tell apart using abundance data alone (McGill *et al.* 2007; Al Hammal *et al.* 2015). These issues are as problematic for the independent species approach as for the zero-sum approach.

Therefore, the independent species sampling formulas must not be applied blindly, but should be combined with techniques to evaluate the reliability of the maximum-likelihood estimates. When applying the sampling formulas in practice, it is important to assess the estimation bias of the model parameters. A common approach consists in simulating many times the community model with the estimated parameter values, and determining the maximum-likelihood parameters for each of the simulated datasets, which are then compared to the simulation values. The zero-sum and independent species model variants present the same parameter estimation biases. However, the evaluation of these biases is more efficient for independent species models, because they are particularly easy to simulate. Simulated datasets are also used to test whether the fitted model can satisfactorily reproduce the empirical data (Etienne 2007; Jabot & Chave 2011).

The flexibility of the independent species assumption allows us to construct new hypothesis tests on a wide range of community processes. However, the reliability of such tests should be carefully assessed. For example, we repeatedly used the tropical forest data to illustrate our sampling formulas. Each of these sampling formulas deals with one or two community processes (including dispersal limitation, different speciation mechanisms, and density dependence), and we determined for each process separately whether it is supported by the data (using Akaike information criterion). A more satisfying approach would combine these processes in a single, nested model, and test whether particular instances of this general model provide fits of similar quality. However, this approach would most

probably lead to overparametrization problems, which can be detected by appropriate model selection techniques (Burnham & Anderson 2003; note that these techniques are often simulation-based). Clearly, the technical possibility to evaluate the likelihood function does not at all guarantee the reliability of the inference results.

Species abundance distributions are known to contain limited information about the processes that structured the community (McGill *et al*. 2007). More powerful inferences might be possible based on abundance data coming from multiple sites, which can be handled with the approach presented in this paper. A similar approach can be instrumental to integrate also other types of data, such as species-area relationships (O'Dwyer & Green 2010), time-series data (Kalyuzhny, Kadmon & Shnerb 2015) and phylogenetic information (Manceau, Lambert & Morlon 2015). Combining different patterns will yield stronger tests of the adequacy of a model to fit the data. To tackle this, the independent species approach seems a promising tool.

Authors' contributions

B.H. and R.S.E. conceived the study, developed the theory, analysed the examples, programmed the R package, and wrote the paper.

Acknowledgements

We thank four anonymous reviewers and S. Dray for insightful comments, and the Center for Tropical Forest Science for data collection. Financial support was provided by the French National Research Agency (ANR) through the TULIP Laboratory of Excellence (to B.H., grant number ANR-10-LABX-41), by the Netherlands Organization for Scientific Research (NWO) (VICI grant number 865.13.003) through VIDI and VICI grants to R.S.E., and by the bilateral French-Dutch Van Gogh programme (to B.H. and R.S.E.).

Data accessibility

All datasets and models analysed in this paper are available in the R package SADISA, which can be downloaded at https://CRAN.R-project.org/package=SADISA.

References

Al Hammal, O., Alonso, D., Etienne, R.S. & Cornell, S.J. (2015) When can species abundance data reveal non-neutrality? *PLoS Computational Biology*, **11**, e1004134.

Allouche, O. & Kadmon, R. (2009) A general framework for neutral models of community dynamics. *Ecology Letters*, **12**, 1287–1297.

Alonso, D. & McKane, A.J. (2004) Sampling Hubbell's neutral theory of biodiversity. *Ecology Letters*, **7**, 901–910.

Burnham, K.P. & Anderson, D.R. (2003) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, New York, NY, USA.

Chisholm, R.A. & Pacala, S.W. (2010) Niche and neutral models predict asymptotically equivalent species abundance distributions in high-diversity ecological communities. *Proceedings of the National Academy of Sciences USA*, **107**, 15821–15825.

Connolly, S.R., Hughes, T.P. & Bellwood, D.R. (2017) A unified model explains commonness and rarity on coral reefs. *Ecology Letters*, **20**, 477–486.

Du, X., Zhou, S. & Etienne, R.S. (2011) Negative density dependence can offset the effect of species competitive asymmetry: a niche-based mechanism for neutral-like patterns. *Journal of Theoretical Biology*, **278**, 127–134.

Etienne, R.S. (2005) A new sampling formula for neutral biodiversity. *Ecology Letters*, **8**, 253–260.

Etienne, R.S. (2007) A neutral sampling formula for multiple samples and an 'exact' test of neutrality. *Ecology Letters*, **10**, 608–618.

Etienne, R.S. (2009) Improved estimation of neutral model parameters for multiple samples with different degrees of dispersal limitation. *Ecology*, **90**, 847–852.

Etienne, R.S. & Alonso, D. (2005) A dispersal-limited sampling theory for species and alleles. *Ecology Letters*, **8**, 1147–1156.

Etienne, R.S., Alonso, D. & McKane, A.J. (2007a) The zero-sum assumption in neutral biodiversity theory. *Journal of Theoretical Biology*, **248**, 522–536.

Etienne, R.S., Apol, M.E.F., Olff, H. & Weissing, F.J. (2007b) Modes of speciation and the neutral theory of biodiversity. *Oikos*, **116**, 241–258.

Etienne, R.S. & Haegeman, B. (2011) The neutral theory of biodiversity with random fission speciation. *Theoretical Ecology*, **4**, 87–109.

Etienne, R.S., Latimer, A.M., Silander, J.A. & Cowling, R.M. (2006) Comment on 'neutral ecological theory reveals isolation and rapid speciation in a biodiversity hot spot'. *Science*, **311**, 610b.

Fisher, R.A., Corbet, A.S. & Williams, C.B. (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology*, **12**, 42–58.

Green, J.L. & Plotkin, J.B. (2007) A statistical theory for sampling species abundances. *Ecology Letters*, **10**, 1037–1045.

Haegeman, B. & Etienne, R.S. (2008) Relaxing the zero-sum assumption in neutral biodiversity theory. *Journal of Theoretical Biology*, **252**, 288–294.

Haegeman, B. & Etienne, R.S. (2010) Self-consistent approach for neutral community models with speciation. *Physical Review E*, **81**, 031911.

Haegeman, B. & Etienne, R.S. (2011) Independent species in independent niches behaveneutrally. *Oikos*, **120**, 961–963.

Harris, K., Parsons, T.L., Ijaz, U.Z., Lahti, L., Holmes, I. & Quince, C. (2017) Linking statistical and ecological theory: Hubbell's unified neutral theory of biodiversity as a hierarchical Dirichlet process. *Proceedings of the IEEE*, **105**, 516–529.

Hubbell, S.P. (2001) *The Unified Neutral Theory of Biodiversity and Biogeography*. vol. 32 of Monographs in Population Biology. Princeton University Press, Princeton, NJ, USA.

Jabot, F. & Chave, J. (2011) Analyzing tropical forest tree species abundance distributions using a nonneutral model and through approximate Bayesian inference. *American Naturalist*, **178**, E37–E47.

Janzen, T., Haegeman, B. & Etienne, R.S. (2015) A sampling formula for ecological communities with multiple dispersal syndromes. *Journal of Theoretical Biology*, **374**, 94–106.

Kalyuzhny, M., Kadmon, R. & Shnerb, N.M. (2015) A neutral theory with environmental stochasticity explains static and dynamic properties of ecological communities. *Ecology Letters*, **18**, 572–580.

MacArthur, R.H. (1957) On the relative abundance of bird species. *Proceedings of the National Academy of Sciences USA*, **43**, 293–295.

Manceau, M., Lambert, A. & Morlon, H. (2015) Phylogenies support out-of-equilibrium models of biodiversity. *Ecology Letters*, **18**, 347–356.

May, F., Huth, A. & Wiegand, T. (2015) Moving beyond abundance distributions: neutral theory and spatial patterns in a tropical forest. *Proceedings of the Royal Society B: Biological Sciences*, **282**, 20141657.

McGill, B.J., Etienne, R.S., Gray, J.S. *et al*. (2007) Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters*, **10**, 995–1015.

Munoz, F., Couteron, P., Ramesh, B.R. & Etienne, R.S. (2007) Estimating parameters of neutral communities: from one single large to several small samples. *Ecology*, **88**, 2482–2488.

O'Dwyer, J.P. & Green, J.L. (2010) Field theory for biogeography: a spatially explicit model for predicting patterns of biodiversity. *Ecology Letters*, **13**, 87–95.

Preston, F.W. (1948) The commonness, and rarity, of species. *Ecology*, **29**, 254–283.

Purves, D.W. & Pacala, S.W. (2005) Ecological drift in niche-structured communities: neutral pattern does not imply neutral process. *Biotic Interactions in the Tropics* (eds D. Burslem, M. Pinard & S. Hartley), pp. 107–138. Cambridge University Press, Cambridge, UK.

Rosindell, J., Cornell, S.J., Hubbell, S.P. & Etienne, R.S. (2010) Protracted speciation revitalizes the neutral theory of biodiversity. *Ecology Letters*, **13**, 716–727.

Rosindell, J., Hubbell, S.P. & Etienne, R.S. (2011) The unified neutral theory of biodiversity and biogeography at age ten. *Trends in Ecology and Evolution*, **26**, 340–348.

Steele, M.A. & Forrester, G.E. (2005) Small-scale field experiments accurately scale up to predict density dependence in reef fish populations at large scales. *Proceedings of the National Academy of Sciences USA*, **102**, 13513–13516.

Volkov, I., Banavar, J.R., He, F., Hubbell, S.P. & Maritan, A. (2005) Density dependence explains tree species abundance and diversity in tropical forests. *Nature*, **438**, 658–661.
Volkov, I., Banavar, J.R., Hubbell, S.P. & Maritan, A. (2003) Neutral theory and relative species abundance in ecology. *Nature*, **424**, 1035–1037.
Volkov, I., Banavar, J.R., Hubbell, S.P. & Maritan, A. (2007) Patterns of relative species abundance in rainforests and coral reefs. *Nature*, **450**, 45–49.
Walker, S.C. (2007) When and why do non-neutral metacommunities appear neutral?. *Theoretical Population Biology*, **71**, 318–331.
Zhou, S.R. & Zhang, D.Y. (2008) A nearly neutral model of biodiversity. *Ecology*, **89**, 248–258.

Received 13 February 2016; accepted 21 April 2017
Handling Editor: Stéphane Dray

Supporting Information

Details of electronic Supporting Information are provided below.

Appendix S1. Poisson distributed abundance frequencies.

Appendix S2. Solution of local community model.

Appendix S3. Solution of metacommunity model.

Appendix S4. Conditioning on local community size.

Appendix S5. Density dependence in metacommunity.

Appendix S6. Density dependence in local community.

Appendix S7. Model with species-specific parameters.

Appendix S8. Summary of sampling formulas.

Table S1. Fits for model with local density dependence.

Table S2. Fits for protracted speciation model.

Table S3. Independent-species fits for a large dataset.